IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 5, 2019, accepted April 21, 2019, date of publication April 30, 2019, date of current version June 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914168

A Bootstrapping Approach With CRF and Deep Learning Models for Improving the Biomedical Named Entity Recognition in Multi-Domains

JUAE KIM[1], YOUNGJOONG KO[2], AND JUNGYUN SEO[1], (Member, IEEE)
[1]Department of Computer engineering, Sogang University, Seoul 04107, South Korea
[2]Department of Computer engineering, Dong-A University, Busan 604-714, South Korea

Corresponding author: Youngjoong Ko (youngjoong.ko@gmail.com)

This work was supported by the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (Development of Automatic Text Data Labeling and Verification Tools for Machine Learning) under Grant 2017-0-00550.

ABSTRACT Biomedical named entity recognition (biomedical NER) is a core component to build biomedical text processing systems, such as biomedical information retrieval and question answering systems. Recently, many studies based on machine learning have been developed for a biomedical NER. The machine learning-based approaches generally require significant amounts of annotated corpora to achieve high performance. However, it is expensive to manually create a large number of high-quality corpora due to the demand for biomedical experts. In addition, most existing corpora have focused on several specific sub-domains, such as disease, protein, and species. It is difficult for a biomedical NER system trained with these corpora to provide much information for biomedical text processing systems. In this paper, we propose a method for automatically generating the machine-labeled biomedical NER corpus that covers various sub-domains by using proper categories from the semantic groups of a unified medical language system (UMLS). We use a bootstrapping approach with a small amount of manually annotated corpus to automatically generate a significant amount of corpus and then construct a biomedical NER system trained with the machine-labeled corpus. At last, we train two machine learning-based classifiers, conditional random fields (CRFs) and long short-term memory (LSTM), with the machine-labeled data to improve performance. The experimental results show that the proposed method is effective to improve performance. As a result, the proposed one obtains higher performance in 23.69% than the model that trained only a small amount of manually annotated corpus in F1-score.

INDEX TERMS Biomedical named entity recognition, bootstrapping, information extraction, semi-supervised learning.

I. INTRODUCTION

An increasing number of studies have been conducted using bioinformatics with natural language processing. Significant amounts of data are currently available in medical domains, making it increasingly important to be able to extract and retrieve high-quality information from this data. As an example, the MEDLINE literature database contains over 24 million abstracts of biomedical journals, and many new abstracts have been added so far. In particular, extracting biomedical terms is one of the core tasks in the analysis of biomedical texts. Biomedical named entity recognition(biomedical NER) task is defined as the identification of biomedical entities from the biomedical texts and their classification into categories such as disease, gene, protein, and drug. Since biomedical terms and their categories play an important role in many tasks of bioinformatics [1] such as relation extraction [2],[3], information retrieval [4], and question answering systems [5], many researchers have developed various methods for correctly extracting biomedical NEs. Rule-based approaches have been typically used to extract a biomedical NE [6],[7], while machine-learning based approaches have recently gained attention. The machine-learning approaches attempt to solve

The associate editor coordinating the review of this manuscript and approving it for publication was Shubhajit Roy Chowdhury.

the problems of rule-based and dictionary-based approaches to recognize new NEs and spelling variations. Shen et al. recognized gene and protein NEs by applying a Hidden Markov Model [8]. Other studies presented a biomedical NER system for detecting protein, DNA, RNA, cell-line, and cell-type entity classes by using Conditional Random Field (CRF) with a variety of traditional and novel features [9],[10]. Leaman et al. suggested a chemical NER system that combines two independent CRF models [11]. Li et al. used deep learning techniques for NER, and these studies are based on the RNN and LSTM [12],[13].

However, most of the previous studies attempted to recognize proteins or genes as biomedical NEs and have focused on several specific sub-types such as disease, cell line, cell type, and species. Thus NE categories were limited to specific sub-domains of biomedicine in each corpus. It can arouse a problem that most of the existing biomedical NE corpora do not sufficiently reflect the whole spectrum of biomedical NEs. For example, GENIA [14], GENETAG [15], and PennBioIE [16] corpora cover gene and protein sub-domains; BioText [17], SCAI Disease [18], and Arizona Disease [19] corpora contain only disease names; and the OrganismTagger [20] and Linnaeus [21] corpora are used to recognize and identify species names. These corpora are too narrowly scoped for applications in a large variety of other biomedical text mining themes, and thus, CALBC [22], CRAFT [23], and i2b2 2010 [24] were developed to cover a wider sub-domain. The CALBC corpus includes proteins and genes, chemicals, diseases and disorders, and living beings and the CRAFT corpus can cover NCBI taxonomy that contains organisms as well as the taxonomy of gene, protein and cell type. The CALBC and the CRAFT contain more sub-domains than other existing corpora but still does not cover many sub-domains such as anatomy, organ, and medical procedures, which are important information in biomedical domains. The i2b2 2010 corpus contains domains with greatly expanded coverage unlike other corpora, and biomedical NEs of the i2b2 2010 corpus contains only three categories including problem, test, and treatment. Because the i2b2 2010 corpus was annotated with phrases referring to very wide sub-domains, biomedical NER system learned this corpus does not represent specific information. For example, the NEs of disease class and symptom class contained in problem class despite disease and symptom are slightly different. Therefore, to develop a useful biomedical NER system for other biomedical tasks, a corpus is required to cover various sub-domains and to represent detail information for each sub-domain as proper categories.

In this paper, we introduce how to obtain a biomedical NE corpus that covers various sub-domains with specific information. It is difficult to manually generate a large-scale and high-quality corpus because manual annotation is extremely time-consuming and expensive, and medical experts are required. For providing sufficient biomedical information with the biomedical NE corpus, we defined the NE category as the semantic group of Unified Medical Language System (UMLS) [25]. The UMLS includes the meta-thesaurus, semantic network, specialist lexicon, and lexical tools. The meta-thesaurus of UMLS is a large biomedical thesaurus that is organized by concepts or meanings, and it links similar names for the same concept from nearly 200 different vocabularies. UMLS semantic types are composed of semantic groups with concepts [26]. There are 133 UMLS semantic types that can be mapped to 15 groups known as UMLS semantic groups[1] which cover various fields in the biomedical domain. The machine-labeled corpus is improved by applying a bootstrapping approach. An effective biomedical NER system can be constructed by learning the machine-labeled NE corpus as training data. In summary, there are three contributions to this paper.

- We propose a method to generate a biomedical NE corpus that covers various biomedical sub-domains by applying UMLS semantic groups as categories of biomedical NE.
- We propose the bootstrapping approach to generate a significant amount of machine-labeled corpus with little human effort.
- Our proposed biomedical NER system makes a higher performance than the existing tool, MetaMap, which uses UMLS information.

As a result, the proposed biomedical NER system that is trained with the machine-labeled corpus by using a bootstrapping approach can provide more specific category information (15 UMLS semantic groups[26]) than NER systems that trained with the other existing corpora mentioned above. The overall paper consists of 4 chapters, including this introductory chapter. Chapter 2 begins an overview of our proposed method, and its subsections show how our proposed model works in detail. In this subsection, we suggest the method of initial corpus generation, bootstrapping approach, machine learning algorithm, and additional feature to enhance the performance of biomedical NER. The third chapter analyses the experimental results using the proposed NER system and the baseline system. In the final chapter, we present a brief summary and future works.

II. PROPOSED METHOD

Figure 1 shows an overview of the proposed method. We will explain the detailed model with three parts. In the first part, marked with circle digit 1 in Figure 1, indicates the method for generating an initial corpus with a chunker and MetaMap. The second part is bootstrapping to improve the quality of the initial corpus. Finally, the bootstrapped initial corpus is used as training data for biomedical NER with CRF and LSTM networks.

A. HOW TO AUTOMATICALLY GENERATE AN INITIAL CORPUS

In this section, we present how to generate an initial corpus to build an initial classifier. Generally, a bootstrapping approach

[1] https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

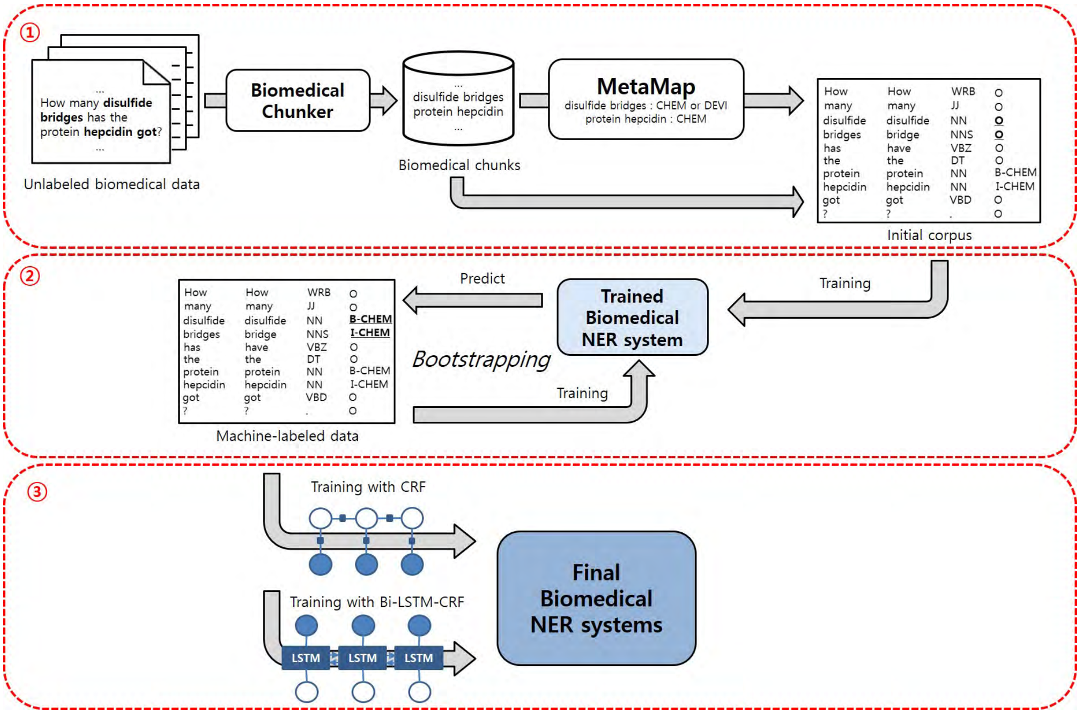


FIGURE 1. **The overview of the proposed method.**

starts with a classifier trained with the initial corpus which is manually annotated (also called seed corpus) and then gradually improves the accuracy of the classifier through several re-training processes. Thus it is important to obtain the high quality of the initial corpus because the performance of the bootstrapping approach seriously depends on the training data for the initial classifier [27],[28]. However, it is sometimes hard to manually annotate sufficient amounts of the initial corpus. Therefore we try to automatically generate a sufficient amount of initial corpus with the chunker and MetaMap.

1) WHY DOES NOT USE METAMAP ALONE FOR AUTOMATIC LABELING?

The reason for using the chunker and MetaMap together is as follows. MetaMap is a useful tool to extract biomedical NEs and match their groups known as concepts to the UMLS meta-thesaurus [29]. Sometimes, MetaMap has been used as a biomedical NER system for various applications[30],[31]. However, there are several limitations to the usage of MetaMap alone as a biomedical NER system [32]: (i) MetaMap extracts not only biomedical NEs but also common entities or even verbs that are not clearly biomedical NEs because the UMLS meta-thesaurus includes a significant amount of common knowledge; (ii) if a biomedical NE does not exist in the UMLS thesaurus, MetaMap cannot determine the UMLS semantic type of the biomedical NE. In other words, it is hard to extract newly generated biomedical NEs with MetaMap. In addition, several previous studies showed that MetaMap performed poorly in their experiments [33],[34]. Thus, it is difficult to obtain a high quality of the initial corpus by using only MetaMap. Therefore, we focus on how to improve the quality of the initial corpus by overcoming the limitations listed above and by exploiting the advantages of MetaMap to enhance the biomedical NER.

2) EXTRACTION OF BIOMEDICAL NE CHUNKS

The chunker was built up to extract only biomedical NE chunks by using a manually annotated corpus, it solves the problem of extracting common entities in the initial corpus. Besides, the chunker enables a term sequence to correspond to a biomedical NE in a sentence by peripheral information associated with the term sequence, even if the term sequence is not registered in UMLS.

We used the features which are listed in Table 1 and word tokens in a window of size two around the current word. We use CoNLL 2003 dataset format proposed by [35] as training data for chunker. That format represents a single word with a series of words, tab-separated features, and label of a word. The features are represented as string symbol, and its representations of examples are described in the column of 'Expression in the labeled corpus.' Word uni-gram, lemma uni-gram, and POS uni-gram are represented as a string of itself in a labeled corpus and the features in row 4-9 in Table 1 are represented as the symbols of 'f1+' to 'f6+.' For example, the word "NXY059" has partial capital letters and Number, so the symbol 'f2+' and 'f3+' annotated as the features of "NXY059." In training data, the example represented as "NXY059 NXY059 NN f2+ f3+

TABLE 1. Feature list for chunker.

Features	Description	Examples	Expression in training corpus
Word unigram	We used the word unigram feature, i.e., a unit of labeling is a word.	How, many, disulfide, bridges, has, the, protein, …	How, many, disulfide, bridges, has, the, protein …
Lemma unigram	We used the lemma of the word.	how, many, disulfide, bridge, have, the protein, …	how, many, disulfide, bridge, have, the protein, …
POS unigram	We used the part-of-speech (POS) of word unigram as the feature.	DT, NN, NNP, RBR, TO, VB, …	DT, NN, NNP, RBR, TO, VB, …
Capital of all letters	If all letters of word are capital, we set this feature.	DNA, RNA, TRH, ACE, TSA, …	f1+
Capital of partial letters	If the word has some capital letters, we set this feature.	Peutz-Jeghers, Marfan, Parkinson disease, …	f2+
Number	If the word has a number, we set this feature.	NXY059, CRISPR-CAS9, …	f3+
Special letter	If the special letters are included in word as ‘*’,‘-’,‘.’, we set this feature.	Li-Fraumeni, CRISPR-CAS9, …	f4+
Capital letters in bracket	If the word is inside the bracket and has a capital letter, we set this feature.	(Tubulin Heterodimers) …	f5+
Quotation mark	If the word exists inside the quotation, we set this feature.	"genomic signatures" …	f6+

B.” Herein, We adopt the IOB2 tagging scheme.[2] The type of label is only three ‘B’, ‘I’ and ‘O’ because the chunker does not determine the category of a word, only extract the range of biomedical NE.

3) AUTOMATIC LABELING FOR THE INITIAL CORPUS

The initial corpus is automatically generated by the chunker and MetaMap with unlabeled medical data. The format of the initial corpus is the same with training data of the chunker, but the initial corpus has different labels which contain category information. We first input the unlabeled medical data into the chunker mentioned the previous section to detect the scope of biomedical NE chunks. The chunking results, biomedical chunks, are used as inputs of MetaMap to determine the UMLS semantic groups as categories of input. The results of MetaMap, UMLS semantic groups, are eventually used to the categories of extracted biomedical NE chunks. to annotate the initial corpus.

However, there are two cases in which a biomedical NE chunk does not have only one UMLS semantic group. First, a biomedical NE chunk may have no UMLS semantic group. If a biomedical NE chunk does not exist in the UMLS, MetaMap cannot analyze the UMLS semantic group of input NE chunks. We refer to this problem as ‘out-of-vocabulary problem’ in this paper. Second, a biomedical NE chunk can have several UMLS semantic groups. A lot of homonyms, especially abbreviation, are existed in the biomedical domain. Thus even the same NE chunks can belong to different UMLS semantic groups. In this case, MetaMap outputs all UMLS semantic groups of biomedical NE chunks can have. This problem is named ‘ambiguity problem.’ In the above two cases, we did not consider those biomedical NE chunks as correct biomedical NEs. Since the labels of the initial corpus were tagged with the IOB2 tagging scheme, biomedical NE chunks that did not have a single UMLS semantic group were annotated with ‘O’ tags in the initial corpus. Therefore, many biomedical NEs cannot be covered by the initial corpus. Consequently, we suggest the method of applying bootstrapping to improve the automatically annotated corpus quality, and it will be discussed in the next section.

B. THE BOOTSTRAPPING TO IMPROVE THE AUTOMATICALLY GENERATED DATA

We present how to apply the bootstrapping approach to our proposed biomedical NER system in this section. In general, a bootstrapping approach starts with a small amount of manually annotated corpus called seed corpus and then generates an initial classifier learned with the seed corpus [36],[37]. After the initial classifier analyzes the unlabeled data, its classification results are added to training data. And then the classifier is re-trained with the machine-labeled corpus and seed corpus. As this process is iterative, the performance of the classifier is gradually increased.

A bootstrapping approach is designed and applied to make the initial corpus more accurate by resolving the problem of incorrect labels with O tags caused by an out-of-vocabulary problem and an ambiguity problem. Algorithm 1 describes the main structure of our bootstrapping approach. In line 1, the initial classifier trained with the initial corpus and the manually labeled seed corpus. It is a little different from previous studies [36],[37] in that our bootstrapping approach starts with the initial corpus which machine-labeled

[2] IOB2 format is tagging format for tagging tokens such as chunking, part-of-speech tagging, NER. There are three kinds of tags in IOB2 format; B-class, I-class and O. The ‘B-class’ indicates that the tag is the beginning of a chunk. An ‘I-class’ indicates that the tag is inside a chunk and ‘O tag’ indicates that a token is not a chunk.

automatically and the manually labeled seed corpus together, not just the seed corpus. By this way, the initial classifier can have more information on a large amount of context from unlabeled medical data and the UMLS thesaurus information by MetaMap.

Algorithm 1 Outline of the Bootstrapping Approach

```
Require:
    $I$ : Initial corpus.
    $U$ : Unlabeled medical data.
    $S$ : Seed corpus.
    $M_i$ : Machine-labeled corpus generated in the $i$th
        bootstrap iteration.
    $C_i$ : Classifier trained in the $i$th bootstrap iteration.
    $Iter_{max}$ : Number of iterations.

 1: Initial classifier $C_0 \leftarrow$ CRF based Classifier $(I, S)$
 2: $t \leftarrow 1$
 3: $M_0 \leftarrow I$
 4: while $t < Iter_{max}$ do
 5:     $M_t \leftarrow$ Annotate on $U$ by $C_{t-1}$
 6:     while $w_i \subseteq U$ do
 7:         $l_i \leftarrow$ the label of $w_i$
 8:         if $l_i$ in $M_{t-1} \neq l_i$ in $M_t$ then
 9:             if $l_i$ in $M_{t-1}$ is 'O' label then
10:                 Tagging $l_i$ in $M_t$ as $l_i$ in $M_t$
11:             else
12:                 Tagging $l_i$ in $M_t$ as $l_i$ in $M_{t-1}$
13:             end if
14:         end if
15:         $C_t \leftarrow$ CRF based Classifier $(M_t, S)$
16:         t $\leftarrow$ t+1
17:     end while
18: end while
19: return $M_{1,\ldots,t}$
```

For the first iteration of bootstrapping, the unlabeled medical data is labeled by the initial classifier (line 5). The first machine-labeled corpus is generated with the results of the initial classifier, by applying followings; in the first criterion, denoted in line 9-10, if any entity with an O tag in the initial corpus is changed to a B or I tag by the classifier, the tag is replaced with the new B, or I tag. The second criterion is denoted in line 11-12. If any entity with the B or I tag in the initial corpus is annotated with the O tag or B or I tag with the different category name, the label of this entity is maintained the same one. As this process is repeated, the O tags in a previous machine-labeled corpus are replaced with correct tags by using the peripheral information. Since we focused on solving the out-of-vocabulary problem and the ambiguity problem, in which biomedical NEs are tagged with O tags in our approach, our method of only replacing O tags into correct labels can generate a more robust machine-labeled corpus.

Figure 2 illustrates our schematized bootstrapping process with the improved machine-labeled corpus and the seed corpus. Short red lines in both of the initial and improved corpora denote biomedical NEs. After several bootstrap iterations, the improved machine-labeled corpus has more accurate biomedical NEs than initial corpus.

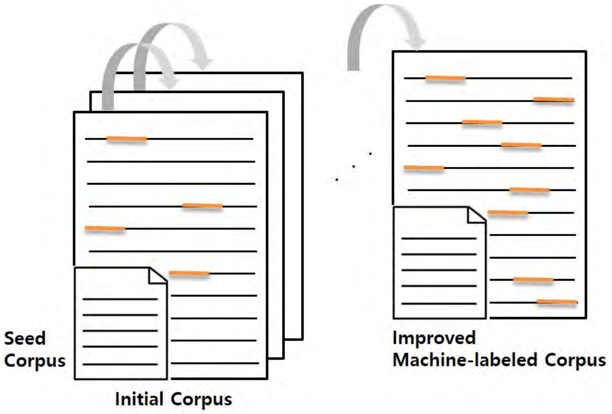
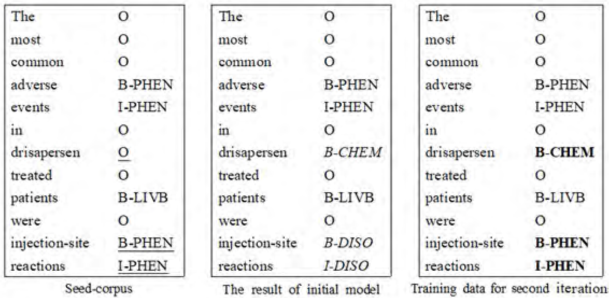


FIGURE 2. Proposed bootstrapping process in terms of corpus.



The	O
most	O
common	O
adverse	B-PHEN
events	I-PHEN
in	O
drisapersen	O
treated	O
patients	B-LIVB
were	O
injection-site	B-PHEN
reactions	I-PHEN

Seed-corpus

The	O
most	O
common	O
adverse	B-PHEN
events	I-PHEN
in	O
drisapersen	*B-CHEM*
treated	O
patients	B-LIVB
were	O
injection-site	*B-DISO*
reactions	*I-DISO*

The result of initial model

The	O
most	O
common	O
adverse	B-PHEN
events	I-PHEN
in	O
drisapersen	**B-CHEM**
treated	O
patients	B-LIVB
were	O
injection-site	**B-PHEN**
reactions	**I-PHEN**

Training data for second iteration

FIGURE 3. Example of generating the training data for next iteration.

Figure 3 shows the example that how to generate the training data for the second iteration using an initial corpus and the results of the first classifier. The entity 'injection-site reactions' is tagged with 'B-PHEN[3]' and 'I-PHEN' in the initial corpus while the entity is classified to 'B-DISO[4]' and 'I-DISO' in the results of the first classifier. In this case, PHEN tags are still maintained. As another example, the entity 'drisapersen' with an O tag in the initial corpus is newly classified by 'B-CHEM[5]' by the first classifier, the tag of the drisapersen is changed into 'CHEM', which is a chemical category of the UMLS semantic groups and includes drugs, proteins, steroids, vitamins, and others. In actual, the drisapersen is a type of drug. Through this process, the label of the drisapersen is corrected in the training data for the second iteration.

[3] This is an abbreviation of phenomena that contains biologic function, human-caused phenomenon or process, and so on.

[4] This is an abbreviation of disorders that contains cell or molecular dysfunction, disease or syndrome, and so on.

[5] This is an abbreviation of chemicals & drugs that contains carbohydrate, chemical, clinical drug, enzyme, lipid, and so on.

C. MACHINE LEARNING BASED BIOMEDICAL NER SYSTEM

This section describes two training methods for implementing the biomedical NER and the feature which improves performance.

1) CRF BASED BIOMEDICAL NER

CRF is a sequence modeling framework, conditional probability distributions on an undirected graph model, for building probabilistic models [38]. We apply linear-chain CRF for our proposed CRF based biomedical NER. The conditional probability of linear-chain CRF determined on observations x and random variables y as follows:

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{j}\sum_{i=1}^{n}\lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_{k}\sum_{i=1}^{n}\mu_k s_k(y_i, x, i)) \quad (1)$$

$Z(x)$ is a scale factor that guarantees that the posterior probabilities sum to one. $t_j(y_{i-1}, y_i, x, i)$ is a transition function of $i-1$ and ith label sequence and all observation sequences. $s_k(y_i, x, i)$is a state feature function of ith observation sequence. λ_j and μ_k are variables that will be estimated from training data with cross-entropy method.

2) BIDIRECTIONAL LSTM-CRF NETWORK BASED BIOMEDICAL NER

We apply a deep learning approach to improve the performance of our biomedical NER system. Recent studies [39]-[42] using deep learning techniques have demonstrated the high performance for sequential labeling tasks contain the NER task. In particular, LSTM-CRF is known as a model with high performance in NER tasks [43],[44]. It shows good performance in a biomedical domain as well [45],[46]. Huang et al. [47] showed that bidirectional LSTM-CRF performed well for sequence tagging tasks such as chunking, named entity recognition system and part-of-speech tagging. To improve performance, we applied bidirectional LSTM-CRF [44], a deep learning-based model, to learn with the final generated corpus. The bidirectional LSTM-CRF model was not used throughout the bootstrapping process because it is a complex model and it requires too long training time for learning. Figure 4 shows the architecture of the bidirectional LSTM-CRF model for our biomedical NER system. The input layer represents input features that consist of word vector representations and one-hot representation of handcrafted features shown in Table 1. We utilize a 3.5GB word embedding provided by BioASQ to convert the input word to a vector representation. The output layer represents the probability distribution over labels at time t. The output dimensionality was 31 that is the number of labels with O tag and the combination of categories and B or I tags. Our bidirectional LSTM-CRF model has 100-dimensional LSTM cells. We train our NER model by using the backpropagation algorithm, updating the weights on every 30 sentences of training data, and using the early stopping criteria. The dropout rate for training is 0.5.

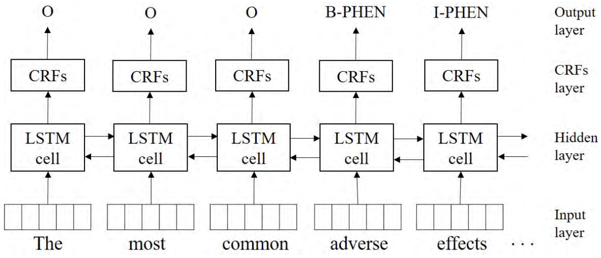


FIGURE 4. Bidirectional LSTM-CRF for named entity recognition.

3) CORPUS-BASED FEATURE

In this section, we introduce a dictionary and context-based feature to improve performance of biomedical NER system. Most of the recognition studies [48],[49] have shown that the dictionary matching was an important feature to improve performance of NER. In the biomedical domain, there is a high-quality of medical dictionary known as UMLS developed by U.S. National Library of Medicine. This dictionary has a vocabulary database of biomedical concepts, their semantic types, and relationship, and it has been used for mapping the contents of a biomedical text to concepts by MetaMap. Thus we try to construct the corpus-based features by using the information of UMLS dictionary from MetaMap.

Herein, we explain how to extract the corpus-based features by the following steps. First, we input the unlabeled data into a chunker and the results of the chunker are tagged by MetaMap. When raw sentences are entered into MetaMap without a chunker, a lot of common entities or verbs are matched with UMLS vocabularies from MetaMap. That is a reason why we utilize the results of a chunker, not only use UMLS vocabularies. The extracted results from this process are called NE features and this process is similar to the process of generating the initial corpus. Then the NE features that have only one semantic group type, no ambiguity, are listed as entries in the corpus-based feature set. This set is referred to as the corpus-based NE set (CBNES).

For using the CBNES as a good feature, the noise from automatic extraction have to be removed. When we observed the entries of CBNES, we found that the entries, NE features, of CBNES with too high or low frequency have high probabilities to be noise. The noisy NE features because, in many cases, NE features with too high-frequency common words are common words that are incorrectly extracted by NE features and ones with too low frequency are a kind of error. Therefore, we set up the removal percentages as two thresholds to get rid of NE features with too high or low frequencies from CBNES. To set up the parameters, we conducted close tests and their results are shown in Figure 5 and 6. Figure 5 illustrates performance changes after getting rid of NE features with low frequencies from 1 to 10 and Figure 6 does

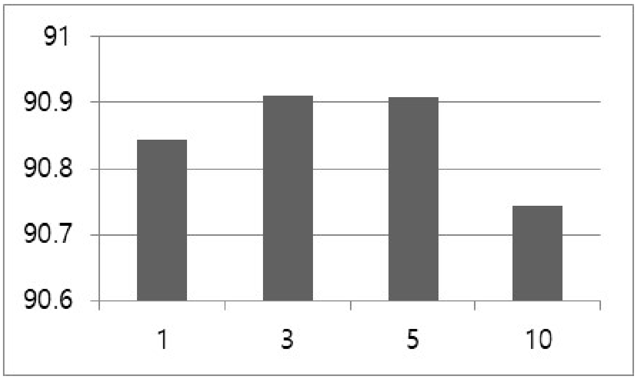


FIGURE 5. Performance changes according to each threshold value that indicates frequencies of the NE features in CBNES.

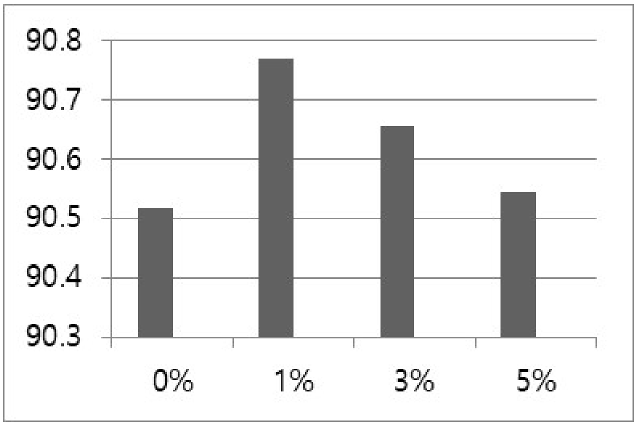


FIGURE 6. Performance changes according to removal percentages of the NE features with top low frequencies in CBNES.

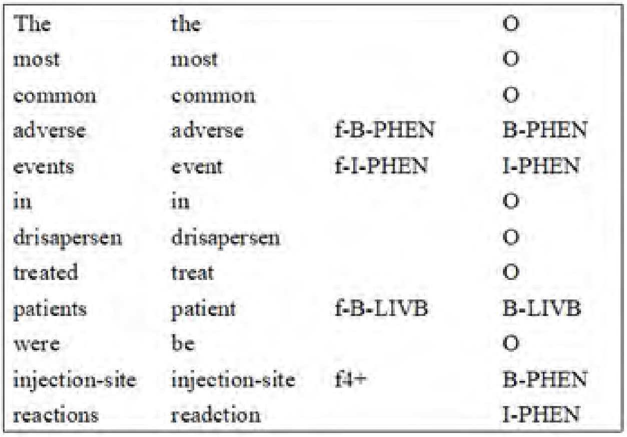
performance changes after getting rid of NE features with the top few percentages from 0% to 5%. According to results of these experiments, extracted NE features located higher than top 1% or with lower than 5 frequencies are removed from CBNES after all NE features in CBNES are sorted in decreasing order by their frequencies.

We illustrate an example of how to express the corpus-based features in the training data through Figure 7. The first column is the word feature and the second column is the lemma feature. In the third column, 'f-B-PHEN' and 'f-B-LIVB' are corpus-based features. It means the entity 'adverse event' is included in CBNES and it has 'PHEN' label. Since the word 'injection-site' has a '-' letter, the special-letter feature 'f4+', which is mentioned in Table 1, is represented as the feature of 'injection-site.' The Final column is the correct answer tag of the word positioned in the first column.

III. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

In our studies, we used two corpora, a small amount of manually annotated corpus and a significant amount of unlabeled data. The manually annotated corpus (MAC) consists of biomedical questions distributed by BioASQ 2015 and 2016 without duplicate questions. The half of the manually annotated corpus was used as test data and the other half was used as seed data. A significant amount of unlabeled data is composed of 7,631 PubMed article abstracts that were arbitrarily selected. Table 2 shows the statistics of these corpora.



The	the		O
most	most		O
common	common		O
adverse	adverse	f-B-PHEN	B-PHEN
events	event	f-I-PHEN	I-PHEN
in	in		O
drisapersen	drisapersen		O
treated	treat		O
patients	patient	f-B-LIVB	B-LIVB
were	be		O
injection-site	injection-site	f4+	B-PHEN
reactions	readction		I-PHEN

FIGURE 7. Training data sample.

TABLE 2. Dataset statistics.

	Manually Annotated Corpus (MAC)			Unlabeled Data
	Training data	Test data	Overall	
Sentences	625	624	1,249	67,925
Words	7,062	7,005	14,067	1,767,498
Entities	1,535	1,492	3,029	-

TABLE 3. Comparing performances of MetaMap as baseline, the CRF model trained with MAC training data and initial classifier with initial training data.

Model	Correct #	Predict #	Precision	Recall	F1-score
MetaMap	878	2,510	34.98%	58.8%	43.88%
CRF with MAC	631	921	68.51%	42.3%	52.3%
Initial classifier	860	1,036	83.01%	57.64%	68.04%

The evaluation metrics to measure the performance of biomedical NER system are precision, recall, and F1. We describe some terminologies for describing our evaluation metrics. 'Correct' is the number of the biomedical NEs that match the correct answer and the count of 'Correct' is measured with the exact matching of the biomedical NEs, not a single word. 'Predict' is the number of biomedical NEs predicted by the NER system. 'Answer' is the number of golden-standard biomedical NEs that we have to recognize in the test data. Precision is to measure the quality of predictions based on that system refers to be positive and precision is represented as the ratio of the number of predicted NEs that are correct answers to the number of biomedical NEs predicted from the proposed NER system like equation (2).

TABLE 4. Comparing performances of CF usages according to the size of corpus.

Model	Correct #	Predict #	Precision	Recall	F1-score
Initial classifier	860	1,036	83.01%	57.64%	68.04%
Initial classifier with 10MB-CF	864	1,036	83.40%	57.91%	68.35%
Initial classifier with 20MB-CF	915	1,131	80.90%	61.33%	**69.77%**

TABLE 5. The number of biomedical NEs detected in training data by bootstrap iterations.

Model	The number of detected biomedical NEs
Initial classifier	255,488
Second bootstrap iteration	256,009(+521)
Third bootstrap iteration	256,122(+113)
Fourth bootstrap iteration	256,176(+54)
Fifth bootstrap iteration	256,189(+13)

Recall measures how much the proposed model can capture the actual answers in test data. Recall is calculated as the number of predicted NEs that is correct answer divided by the number of real answers in test data like equation (3). F1-score is defined as equation (4) as the harmonic average of precision and recall.

$$Precision = \frac{\text{num of system predictions}}{\text{num of correct predictions}} = \frac{Predict}{Correct} \quad (2)$$

$$Recall = \frac{\text{num of gold arguments}}{\text{num of correct predictions}} = \frac{Answer}{Correct} \quad (3)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

B. EXPERIMENTAL RESULTS

1) COMPARING THE INITIAL RESULTS OF THE PROPOSED METHOD USING A CHUNKER TO METAMAP

In this section, we show an effect of a machine-labeled corpus. The MetaMap model in Table 3 is a model that used only MetaMap as a biomedical NER system. In this model, the sentences of test data are entered to MetaMap for recognizing the biomedical NEs. The CRF with MAC model is trained with only a small amount of MAC called the seed corpus. The initial classifier model was learned with the training data that consists of the initial machine-labeled corpus and the seed corpus.

Performance of the MetaMap model is lower than the CRF with MAC model. In particular, the precision score of the MetaMap model is much worse than its recall score because the MetaMap model makes a significant amount of false-positive errors. That is, the CRF with MAC model does not extract the common nouns and verbs by learning with the annotated corpus for biomedical NEs, but the MetaMap model extracts many common nouns and verbs for them. In the other hand, the recall score of the CRF with MAC model showed lower performance than MetaMap because the small amount of training data cannot cover various training examples of whole classes such as anatomy, Living Beings, occupations and so on.

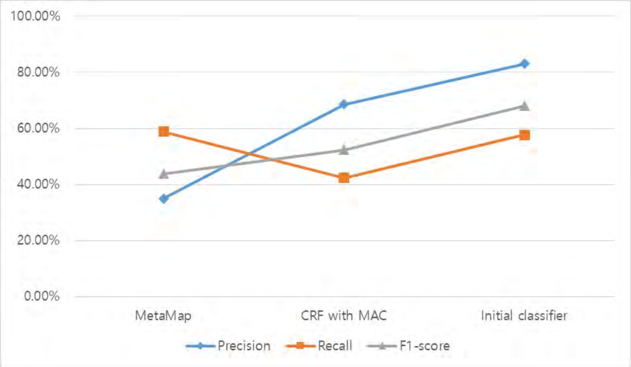


FIGURE 8. Graphical representation of comparing of MetaMap, CRF with MAC and initial classifier.

By comparing the initial classifier with the MAC model, we can verify the impact of a significant amount of machine-labeled corpus. The initial classifier makes a higher performance than the CRF with MAC model with 15.34% increased recall score and F1-score is improved up to 68.04%. Because the training data for the initial classifier has more various examples and information by using a significant amount of training data that can cover UMLS semantic groups as the NE categories. Although the initial classifier cannot recognize some terms that have the out-of-vocabulary problem or the ambiguity problem as biomedical NEs, it showed better performance than MetaMap and the CRF with MAC model. The F1-score of the initial classifier is much higher than the MetaMap model at 24.16%. These results show that a large amount of data was important and the proposed method of generating initial-machine labeled data can improve the performance of biomedical NER system. Figure 8 is a graphical representation of the results corresponding to Table 3.

2) COMPARING THE RESULTS OF CORPUS-BASED FEATURES ACCORDING TO THE SIZE OF CORPUS

In the previous section, we explained corpus-based features to reflect the information of UMLS thesaurus. Experimental results using the corpus-based features without the bootstrapping approach are shown in Table 4. The terms 10MB-CF and 20MB-CF indicate that corpus-based features were extracted from 10MB and 20MB unlabeled data composed of PubMed

TABLE 6. Performance change according to bootstrapping approach with corpus-based feature.

Model	Correct #	Predict #	Precision	Recall	F1-score
MetaMap	878	2,510	34.98%	58.80%	43.88%
CRF with MAC	631	921	68.51%	42.30%	52.3%
Initial classifier with 20MB-CF	915	1,131	80.90%	61.33%	69.77%
Second bootstrap iteration model with 20MB-CF	912	1,091	83.59%	61.16%	70.62%
Third bootstrap iteration model with 20MB-CF	921	1,071	85.99%	61.73%	**71.87%**
Fourth bootstrap iteration model with 20MB-CF	900	1,062	84.75%	60.32%	70.48%
Fifth bootstrap iteration model with 20MB-CF	915	1,083	84.49%	61.33%	71.07%

TABLE 7. Enhancing performance of biomedical NER system using Bidirectional LSTM-CRF.

Model	Correct #	Predict #	Precision	Recall	F1-score
CRF with MAC	631	921	68.51%	42.30%	52.3%
CRF with MAC+MLC	847	1,071	85.99%	61.73%	71.87%
Bidirectional LSTM-CRF with MAC	772	1,232	73.21%	60.46%	66.23%
Bidirectional LSTM-CRF with MAC+MLC	941	1,203	85.12%	68.63%	**75.99%**

articles. As shown in Table 4, corpus-based features from a larger corpus more contributed to improving the biomedical NER system.

3) PERFORMANCES OF BIOMEDICAL NER SYSTEM WITH BOOTSTRAPPING AND CORPUS-BASED FEATURES

Table 5 shows the number of biomedical NEs detected in the machine-labeled corpus for each iteration of the bootstrapping approach. During the five bootstrap iterations, the number of detected biomedical NEs were constantly increased, but the amount of increase was reduced as the bootstrapping iterations progressed. It shows that our bootstrapping approach changes the O tags to other meaningful tags with B or I tags and their entities.

Table 6 shows performance changes in the bootstrapping approach according to the number of iteration with 20MB-CF. It experimentally verifies that our biomedical NER system was improved by overcoming the ambiguity problem and the out-of-vocabulary problem using our proposed bootstrapping approach. However, performances of models in the fourth and fifth iterations were lower than that of the model in the third iteration model. This result shows that too many bootstrapping steps can decrease performance because more wrongly recognized NEs are added to machine-labeled data. The comparison of the initial classifier with 20MB-CF and the third iteration model achieved the best performance can give us evidence for the effect of the bootstrapping method. We improved the F1-score from 69.77% to 71.87% with our bootstrapping method. In addition, the third iteration model outperformed the baseline of MetaMap, with 29.99% improvements.

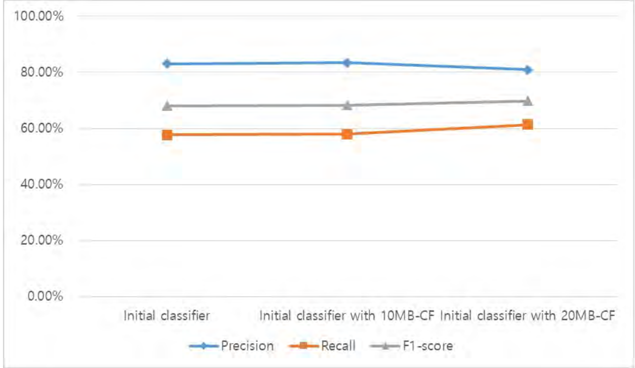


FIGURE 9. Graphical representation of performance changes by the size of corpus for CF.

4) ENHANCING PERFORMANCE OF THE BIOMEDICAL NER SYSTEM WITH DEEP LEARNING

Table 7 shows a comparison of performances of biomedical NER trained by CRF and bidirectional LSTM-CRF with a small amount of MAC and automatically machine-labeled corpus (MLC). As the third bootstrap iteration model with 20MB-CF showed the best performance in Table 6,

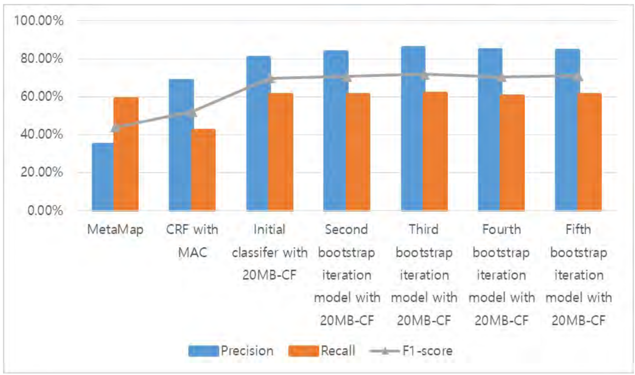


FIGURE 10. Performance changes according to proposed method.

we considered the training data for this model as MLC that is regarded as a final automatically generated corpus using our proposed approach in our following experiments. The experimental results in this subsection show that, regardless of the learning classifiers, the performance of the model learned with the MAC, and MLC data was better than one of the model learned only MAC. It demonstrates that a significant amount of corpus that was automatically generated by our proposed method is effective to improve the performance of biomedical NER system. In particular, deep learning requires a significant amount of training data because a huge number of parameters are required to be tuned by a learning algorithm. Therefore we expect that the proposed method, which automatically generates a significant amount of corpus with the bootstrapping approach, can enhance the performance of other studies using deep learning algorithms. Finally, we increased the F1-score from 71.87% to 75.99% by applying a deep learning algorithm, Bidirectional LSTM-CRF.

IV. CONCLUSION

In this study, we have proposed an effective biomedical NER system that can reduce lots of cost for generating the training data and a problem that a corpus cannot cover various sub-domains with specific information. By applying the UMLS semantic groups as categories of biomedical NEs with MetaMap, we developed a biomedical NER system that provides various and specific information in 15 categories. To generate a significant amount of the training data with a little cost, we proposed the method for automatically and accurately generating the machine-labeled corpus with the bootstrapping approach. In addition, we used a corpus-based feature and bidirectional LSTM-CRF, a deep learning algorithm, to enhance the performance of the biomedical NER. Finally, the proposed system showed 32.11% better performance than MetaMap.

Our proposed method can be useful for many other domains and tasks as well because our approach can construct a high-quality machine-labeled corpus with only a small amount of training data. In particular, generating a significant amount of data can facilitate deep learning based approaches because deep learning algorithms require large training data to achieve high performance in general. In addition, using MetaMap as an open toolkit, developers can build a biomedical NER system without any help from experts in biomedical domains. Unfortunately, our approach does not remove the wrong labels generated during bootstrapping. Therefore we have a plan to apply external resources or various approaches such as lexico-syntactic pattern and bagging as future works. We expect that these approaches will be able to increase the performance of biomedical NER system by removing noises of machine-labeled data, it improves the quality of machine-labeled data.

REFERENCES

[1] C. Yao, Y. Qu, B. Jin, L. Guo, C. Li, W. Cui, and L. Feng, "A convolutional neural network model for online medical guidance," *IEEE Access*, vol. 4, pp. 4094–4103, 2016.

[2] S. C. Onye, A. Akkeleş, and N. Dimililer, "relSCAN–a system for extracting chemical-induced disease relation from biomedical literature," *J. Biomed. Inform.*, vol. 87, pp. 79–87, Nov. 2018.

[3] W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang, and F. Liu, "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinf.*, vol. 19, no. 1, p. 233, Jun. 2018.

[4] T. Groza and K. Verspoor, "Assessing the impact of case sensitivity and term information gain on biomedical concept recognition," *PLoS One*, vol. 10, no. 3, Mar. 2015, Art. no. e0119091.

[5] M. Sarrouti and S. O. E. Alaoui, "A biomedical question answering system in BioASQ 2017," in *Proc. BioNLP*, Vancouver, BC, Canada, Aug. 2017, pp. 296–301.

[6] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "ProMiner: Rule-based protein and gene entity recognition," *BMC Bioinf.*, vol. 6, no. 1, p. S14, May 2005.

[7] S. Sekine and C. Nobata, "Definition, dictionaries and tagger for extended named entity hierarchy," in *Proc. LREC*, Lisbon, Portugal, May 2004, pp. 1977–1980.

[8] D. Shen, J. Zhang, G. Zhou, J. Su, and C.-L. Tan, "Effective adaptation of a hidden Markov model-based named entity recognizer for biomedical domain," in *Proc. ACL Workshop Natural Lang. Process. Biomed.*, Sapporo, Japan, vol. 13, Jul. 2003, pp. 49–56.

[9] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 601–606, Apr. 2011.

[10] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in *Proc. JNLPBA*, Barcelona, Spain, 2004, pp. 104–107.

[11] R. Leaman, C.-H. Wei, and Z. Lu, "tmChem: A high performance approach for chemical named entity recognition and normalization," *J. Cheminf.*, vol. 7, no. 1, p. S3, Jan. 2015.

[12] L. Li, L. Jin, Y. Jiang, and D. Huang, "Recognizing biomedical named entities based on the sentence vector/twin word embeddings conditioned bidirectional LSTM," in *Proc. Int. Symp. Natural Lang. Process. Based Naturally Annotated Big Data China Nat. Conf. Comput. Linguistics*, Yantai, China, Oct. 2016, pp. 165–176.

[13] L. Li, L. Jin, Z. Jiang, D. Song, and D. Huang, "Biomedical named entity recognition based on extended recurrent neural networks," in *Proc. BIBM*, Washington, DC, USA, Nov. 2015, pp. 649–652.

[14] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus–a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. 1, pp. i180–i182, Jul. 2003.

[15] L. Tanabe, N. Xie, L. H. Thom, W. Matten, and W. J. Wilbur, "GENETAG: A tagged corpus for gene/protein named entity recognition," *BMC Bioinf.*, vol. 6, no. 1, p. S3, May 2005.

[16] S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein, and L. Ungar, "Integrated annotation for biomedical information extraction," in *Proc. HLT/NAACL*, Boston, MA, USA, May 2004, pp. 61–68.

[17] B. Rosario and M. A. Hearst, "Classifying semantic relations in bioscience texts," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, Barcelona, Spain, Jul. 2004, Art. no. 430.

[18] H. Gurulingappa, R. Klinger, M. Hofmann-Apitius, and J. Fluck, "An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature," in *Proc. 2nd Workshop Building Evaluating Resour. Biomed. Text Mining*, Valletta, Malta, May 2010, pp. 15–22.

[19] R. Leaman, C. Miller, and G. Gonzalez, "Enabling recognition of diseases in biomedical text with machine learning: Corpus and benchmark," in *Proc. Symp. Lang. Biol. Med.*, Jan. 2009, pp. 82–89.

[20] N. Naderi, T. Kappler, C. J. O. Baker, and R. Witte, "OrganismTagger: Detection, normalization and grounding of organism entities in biomedical documents," *Bioinformatics*, vol. 27, pp. 2721–2729, Oct. 2011.

[21] M. Gerner, G. Nenadic, and C. M. Bergman, "LINNAEUS: A species name identification system for biomedical literature," *BMC Bioinf.*, vol. 11, p. 85, Feb. 2010.

[22] D. Rebholz-Schuhmann, A. J. J. Yepes, E. M. van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, and U. Hahn, "CALBC silver standard corpus," *J. Bioinf. Comput. Biol.*, vol. 8, no. 1, pp. 163–179, Feb. 2010.

[23] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, Jr., K. B. Cohen, K. Verspoor, J. A. B. Blake, and L. E. Hunter, "Concept annotation in the CRAFT corpus," *BMC Bioinf.*, vol. 13, p. 161, Jul. 2012.

[24] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *J. Amer. Med. Informat. Assoc.*, vol. 18, no. 5, pp. 552–556, Jun. 2011.

[25] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, pp. D267–D270, Jan. 2004.

[26] C. P. Morrey, "Mapping of MalaCards maladies to UMLS concepts," in *Proc. BIBM*, Madrid, Spain, Dec. 2018, pp. 1998–2000.

[27] V.-T. Phi, J. Santoso, M. Shimbo, and Y. Matsumoto, "Ranking-based automatic seed selection and noise reduction for weakly supervised relation extraction," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, Jul. 2018, pp. 89–95.

[28] Z. Kozareva and E. Hovy, "Not all seeds are equal: Measuring the quality of text mining seeds," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Los Angeles, CA, USA, Jun. 2010, pp. 618–626.

[29] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program," in *Proc. AMIA Symp.*, Washington, DC, USA, Nov. 2001, pp. 17–21.

[30] N. Collier, A. Oellrich, and T. Groza, "Concept selection for phenotypes and diseases using learn to rank," *J. Biomed. Semantics*, vol. 6, p. 24, Jun. 2015.

[31] R. Reátegui and S. Ratté, "Comparison of MetaMap and cTAKES for entity extraction in clinical notes," *BMC Med. Inform. Decis. Making*, vol. 18, p. 74, Sep. 2018.

[32] A. B. Abacha and P. Zweigenbaum, "Medical entity recognition: A comparison of semantic and statistical methods," in *Proc. BioNLP Workshop*, Portland, OR, USA, Jun. 2011, pp. 56–64.

[33] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *J. Biomed. Inform.*, vol. 46, no. 6, pp. 1088–1098, Dec. 2013.

[34] Y. Kim, E. Riloff, and J. F. Hurdle, "A study of concept extraction across different types of clinical notes," in *Proc. AMIA Annu. Symp.*, Chicago, IL, USA, Nov. 2015, pp. 737–746.

[35] E. F. T. K. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," Jun. 2003, *arXiv:cs/0306050*. [Online]. Available: https://arxiv.org/abs/cs/0306050

[36] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proc. 39th Annu. Meeting Assoc. Comput. Linguistics*, Toulouse, France, Jul. 2001, pp. 26–33.

[37] M. Khordad and R. E. Mercer, "Identifying genotype-phenotype relationships in biomedical text," *J. Biomed. Semantics*, vol. 8, no. 1, p. 57, Dec. 2017.

[38] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, Jun. 2001, pp. 1–10.

[39] W. Khan, A. Daud, K. Khan, J. A. Nasir, M. Basheri, N. Aljohani, and F. S. Alotaibi, "Part of speech tagging in Urdu: Comparison of machine and deep learning approaches," *IEEE Access*, vol. 7, pp. 38918–38936, 2019.

[40] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, Jul. 2017.

[41] J. Hammerton, "Named entity recognition with long short-term memory," in *Proc. 7th Conf. Natural Lang. Learn. HLT-NAACL*, Edmonton, AB, Canada, May 2003, pp. 172–175.

[42] C. N. dos Santos and V. Guimarães, "Boosting named entity recognition with neural character embeddings," May 2015, *arXiv:1505.05008*. [Online]. Available: https://arxiv.org/abs/1505.05008

[43] M. Gridach, "Character-level neural network for biomedical named entity recognition," *J. Biomed. Inform.*, vol. 70, pp. 85–91, Jun. 2017.

[44] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," Mar. 2016, *arXiv:1603.01360*. [Online]. Available: https://arxiv.org/abs/1603.01360

[45] P. Corbett and J. Boyle, "Chemlistem: Chemical named entity recognition using recurrent neural networks," *J. Cheminf.*, vol. 10, p. 59, Dec. 2018.

[46] N. Greenberg, T. Bansal, P. Verga, and A. McCallum, "Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets," in *Proc. EMNLP*, Brussels, Belgium, Oct./Nov. 2018, pp. 2824–2829.

[47] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," Aug. 2015, *arXiv:1508.01991*. [Online]. Available: https://arxiv.org/abs/1508.01991

[48] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. 13th Conf. Comput. Natural Lang. Learn.*, Boulder, CO, USA, Jun. 2009, pp. 147–155.

[49] W. W. Cohen and S. Sarawagi, "Exploiting dictionaries in named entity extraction: Combining semi-Markov extraction processes and data integration methods," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Seattle, WA, USA, Aug. 2004, pp. 89–98.

JUAE KIM was born in Seoul, South Korea, in 1993. She received the B.S. and M.S. degrees in computer engineering from Sogang University, Seoul, in 2015 and 2017, respectively, where she is currently pursuing the Ph.D. degree.

Her research interests include natural language processing, question answering systems, information extraction, and machine learning (include deep learning).

YOUNGJOONG KO received the Ph.D. degree from the Department of Computer Science, Sogang University, Seoul, South Korea, in 2003.

He was with LG-EDS, from 1996 to 1997. Since 2004, he has been with the faculty of Dong-A University, Busan, where he led the Intelligent System Laboratory, Department of Computer Engineering. He was with the Computational Linguistics and Information Processing Laboratory (CLIP), University of Maryland, College Park, as a Visiting Scholar, from 2011 to 2012. His research interests include natural language processing, machine learning (deep neural networks), spoken dialogue systems, information retrieval, and big data analysis.

JUNGYUN SEO received the B.S. degree in mathematics, in 1981, and the M.S. and Ph.D. degrees in computer science from the Department of Computer Science, The University of Texas at Austin, Austin, in 1985 and 1990, respectively.

Since 1991, he has been with the faculty of the Korea Advanced Institute of Science and Technology, Taejon, where he led the Natural Language Processing Laboratory, computer Science Department. In 1995, he moved to Sogang University, Seoul, and became a Full Professor, in 2001. His research interests include multi-modal dialogues, statistical methods for NLP, machine translation, and information retrieval. He served as the President of the Korea Information Science Society, in 2013.